Diffusion Reflectance Map: Single-Image Stochastic Inverse Rendering of Illumination and Reflectance Supplementary Material

Yuto Enyo Ko Nishino
Graduate School of Informatics, Kyoto University
https://vision.ist.i.kyoto-u.ac.jp/

References in this supplementary material refer to the citation numbers in the main text for those citations already made in the paper. For new references, we continue the numbering from the main text.

A. The Reflectance Model

We describe the full reflectance model we use in DRMNet which is based on the Disney principled BRDF model [11]. The model is composed of diffuse reflection f_{diff}, retro reflection f_{retro}, and specular reflection f_{spec}

$$\begin{aligned} &f_r(\boldsymbol{\omega}'_{\text{i}}, \boldsymbol{\omega}'_{\text{o}}; \Psi := \{\boldsymbol{\rho}_{\text{d}}, \rho_{\text{s}}, \alpha, \gamma\}) \\ &= (1-\gamma)\frac{\boldsymbol{\rho}_{\text{d}}}{\pi}(f_{\text{diff}}(\boldsymbol{\omega}'_{\text{i}}, \boldsymbol{\omega}'_{\text{o}}) + f_{\text{retro}}(\omega'_{\text{i}}, \omega'_{\text{o}}; \alpha)) \\ &+ f_{\text{spec}}(\boldsymbol{\omega}'_{\text{i}}, \boldsymbol{\omega}'_{\text{o}}; \boldsymbol{\rho}_{\text{d}}, \rho_{\text{s}}, \alpha, \gamma)\,. \end{aligned} \tag{1}$$

The diffuse reflection f_{diff} depends on the angle of incidence θ_{i} and that of outgoing direction θ_{o}

$$f_{\text{diff}} = \left(1 - \frac{F_{\text{i}}}{2}\right)\left(1 - \frac{F_{\text{o}}}{2}\right), \tag{2}$$

where $F_{\text{i}} = (1-\cos\theta_{\text{i}})^5$ and $F_{\text{o}} = (1-\cos\theta_{\text{o}})^5$. The retro reflection f_{retro} is defined by the metallic parameter γ

$$f_{\text{retro}} = R_R\left(F_{\text{i}} + F_{\text{o}} + F_{\text{i}}F_{\text{o}}(R_R - 1)\right), \tag{3}$$

where $R_R = 2\gamma\cos^2\theta_{\text{d}}$ and θ_{d} is the angle between the half vector $\boldsymbol{h}$ and the incident direction. The specular reflection f_{spec} is based on the microfacet model with the GGX distribution

$$f_{\text{spec}} = \frac{FDG}{4\cos\theta_{\text{i}}\cos\theta_{\text{o}}}, \tag{4}$$

where D is a microfacet distribution function defined by the roughness parameter α

$$D = \frac{\alpha^4}{\pi\left((\boldsymbol{h}\cdot\boldsymbol{n})^2(\alpha^4-1)+1\right)^2}, \tag{5}$$

$\boldsymbol{n}$ is the normal direction of the surface in the local frame, and G is the shadowing-masking function

$$G_1(\boldsymbol{\omega}) = \frac{2}{1+\sqrt{1+\alpha^4(1-\boldsymbol{\omega}\cdot\boldsymbol{n})^2/(\boldsymbol{\omega}\cdot\boldsymbol{n})^2}}, \tag{6}$$

and $G = G_1(\boldsymbol{\omega}'_{\text{i}})G_1(\boldsymbol{\omega}'_{\text{o}})$. F is the Fresnel term which is a combination of

$$F_{\text{dielectric}} = \frac{\left(\dfrac{\cos\theta_{\text{o}} - \eta\cos\theta_{\text{t}}}{\cos\theta_{\text{o}} + \eta\cos\theta_{\text{t}}}\right)^2 + \left(\dfrac{\cos\theta_{\text{t}} - \eta\cos\theta_{\text{o}}}{\cos\theta_{\text{t}} + \eta\cos\theta_{\text{o}}}\right)^2}{2}, \tag{7}$$

where $\eta = \dfrac{2}{1-\sqrt{0.08\rho_{\text{s}}}} - 1$, and

$$F_{\text{Schlick}} = \boldsymbol{\rho}_{\text{d}} + (1-\boldsymbol{\rho}_{\text{d}})(1-\cos\theta_{\text{d}})^5, \tag{8}$$

and

$$F = (1-\gamma)F_{\text{dielectric}} + \gamma F_{\text{Schlick}}. \tag{9}$$

B. Objective Function Derivation

Let us derive the objective function of the reflectance map

$$\mathcal{L}_i = \mathbb{E}_{L_i, f_r, k}|\mu_{\theta,\phi}(\mathbf{n}; L_r^{(k)}, L_r^{(K)}) - L_r(\mathbf{n}; L_i, f_r^{(k)})|_2^2. \tag{10}$$

step by step. The objective of the reverse process is to maximize the marginalized likelihood of the reverse process $p_{\theta,\phi}(L_r^{(0)}|L_r^{(K)})$. Instead of directly maximizing this likelihood, we minimize the negative log likelihood

$$\begin{aligned} &-\log p_{\theta,\phi}(L_r^{(0)}|L_r^{(K)}) \\ &\leq \mathbb{E}_q\left[-\log\frac{p_{\theta,\phi}(L_r^{(0:K-1)}|L_r^{(K)})}{q(L_r^{(1:K-1)}|L_r^{(0)}, L_r^{(K)}, f_r^{(K)})}\right] =: \mathcal{L}_p\,, \end{aligned} \tag{11}$$

Here, $\mathcal{L}_p$ is

$$\begin{aligned}
\mathcal{L}_p &= \mathbb{E}_q[-\sum_{k=2}^{K}\log\frac{p_{\theta,\phi}(L_r^{(k-1)}|L_r^{(k)},L_r^{(K)})}{q(L_r^{(k-1)}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}}^{(K)})} \\
&\quad -\log p_{\theta,\phi}(L_r^{(0)}|L_r^{(1)},L_r^{(K)})] \\
&= \sum_{k=2}^{K}\mathbb{E}_{q_{/\{k-1\}}}D_{KL}[q(L_r^{(k-1)}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}}^{(K)}) \\
&\quad ||p_{\theta,\phi}(L_r^{(k-1)}|L_r^{(k)},L_r^{(K)})] \\
&\quad -\mathbb{E}_q\log p_{\theta,\phi}(L_r^{(0)}|L_r^{(1)},L_r^{(K)}) \\
&= \sum_{k=2}^{K}\mathbb{E}_{q_k}D_{KL}[q(L_r^{(k-1)}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}}^{(K)}) \\
&\quad ||p_{\theta,\phi}(L_r^{(k-1)}|L_r^{(k)},L_r^{(K)})] \\
&\quad -\mathbb{E}_{q_1}\log p_{\theta,\phi}(L_r^{(0)}|L_r^{(1)},L_r^{(K)})\,, \qquad (12)
\end{aligned}$$

where

$$\mathbb{E}_q = \mathbb{E}_{q(L_r^{(1:K-1)}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}}^{(K)})} \qquad (13)$$

$$\mathbb{E}_{q_{/\{k-1\}}} = \mathbb{E}_{q(L_r^{\{1:K-1\}/\{k-1\}}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}}^{(K)})} \qquad (14)$$

$$\mathbb{E}_{q_k} = \mathbb{E}_{q(L_r^{(k)}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}}^{(K)})} \qquad (15)$$

$$\mathbb{E}_{q_1} = \mathbb{E}_{q(L_r^{(1)}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}}^{(K)})} \qquad (16)$$

By modeling the forward process

$$q(L_r^{(k)}|L_r^{(0)},f_r^{(K)}) = \mathcal{N}(L_r^{(k)}|L_r(\mathbf{n};L_i,f_r^{(k)}),\sigma^2\mathbf{I})\,, \qquad (17)$$

and its reverse process

$$p_{\theta,\phi}(L_r^{(k-1)}|L_r^{(k)},L_r^{(K)},f_r^{(k)}) = \mathcal{N}(L_r^{(k-1)}|\mu_{\theta,\phi}(L_r^{(k)},L_r^{(K)}),\delta^2\mathbf{I})\,. \qquad (18)$$

the first term of Eq. (12) becomes

$$\begin{aligned}
&D_{KL}\mathcal{N}(L_r^{(k)}|R(\ell,f_{\mathrm{r}}^{(k)}),\sigma^2\boldsymbol{I}) \\
&||\mathcal{N}(L_r^{(k-1)}|\boldsymbol{\mu}_{\theta,\phi}(L_r^{(k)},L_r^{(K)},\boldsymbol{z}^{(k)}),\delta^2\mathbf{I}) \\
&= \frac{1}{2}[2N\log\frac{\delta}{\sigma} - N + \frac{1}{\sigma^2} \\
&||R(\ell,f_{\mathrm{r}}^{(k)}) - \mu_{\theta,\phi}(L_r^{(k)},L_r^{(K)},\boldsymbol{z}^{(k)})||_2^2 + \frac{\sigma^2}{\delta^2}N]\,, \qquad (19)
\end{aligned}$$

and the second term becomes

$$\begin{aligned}
&\frac{1}{2\delta^2}\mathbb{E}_{q(L_r^{(1)}|L_r^{(0)},L_r^{(K)},f_{\mathrm{r}})}\left[||\mu_{\theta,\phi}(L_r^{(1)},L_r^{(K)},\boldsymbol{z}^{(1)}) - L_r^{(0)}||_2^2\right] \\
&+ N\log(\delta\sqrt{2\pi})\,. \qquad (20)
\end{aligned}$$

By focusing on those terms related to the model parameters θ,ϕ, we obtain the simplified objective in the main text (Eq. (10)).

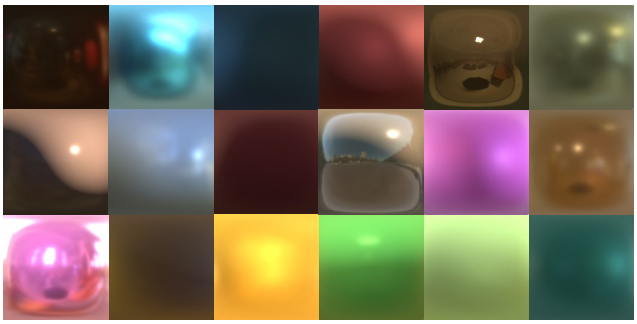

Figure 1. Samples from the synthetic reflectance map dataset we create to train DRMNet. The dataset consists of various combinations of HDR environment maps and reflectance.

C. Network Architectures

As in regular probabilistic diffusion models [30, 87], we use a U-Net with skip connections as the network architecture for IllNet which we denote as ε_θ in $\mu_\theta(L_r^{(k)},L_r^{(K)},\Psi_\phi^{(k)}) = L_r^{(k)} + \varepsilon_\theta(L_r^{(k)},L_r^{(K)},\Psi_\phi^{(k)})$. The input to IllNet is the concatenated observed reflectance map and current reflectance map, $L^{(K)}$ and $L^{(k)}$, respectively. Each layer of the encoder and the decoder contains two residual blocks and consists of 1 to 6 times of 128 channels in increment of 1 from the highest resolution layer. Each residual block additively embeds the current reflectance parameter $\Psi^{(k)}$ to the feature map. In contrast to a regular probabilistic diffusion model which uses sinusoidal positional encoding of the time step together with an MLP to compute the embedding vector, we directly use an MLP to compute the embedding vector from $\Psi^{(k)}$. The input observed reflectance map is 128×128 in resolution. At low-resolution layers of 16×16, 8×8, and 4×4, we apply self-attention to the feature map after the residual block.

For RefNet, we use an encoder of a U-Net with an MLP. As the same as IllNet, the input to RefNet is the concatenated observed reflectance map $L^{(K)}$ and the current reflectance map $L^{(k)}$, and the output is a 6-dimensional vector $\Psi^{(K)}$ of the observed object's reflectance parameter. Each layer of the encoder consists of two residual blocks and has 1, 1, 2, 3, and 4 times of 128 channels from the highest-resolution. The encoder uses the traditional sinusoidal positional encoding and an MLP to additively embed the time steps taken so far $(K-k)$ into each residual block. At the lower-resolution layers of 16×16 and 8×8, we use self-attention.

For ObsNet, we use a U-Net with 1, 2, 3, 4, and 5 times of 128 channels at each layer and self-attention at the 16×16 and 8×8 low-resolution layers. Similar to existing diffusion inpainting methods [60], the network learns to inpaint through inverse diffusion by conditioning on the sparse raw observed reflectance map with missing regions

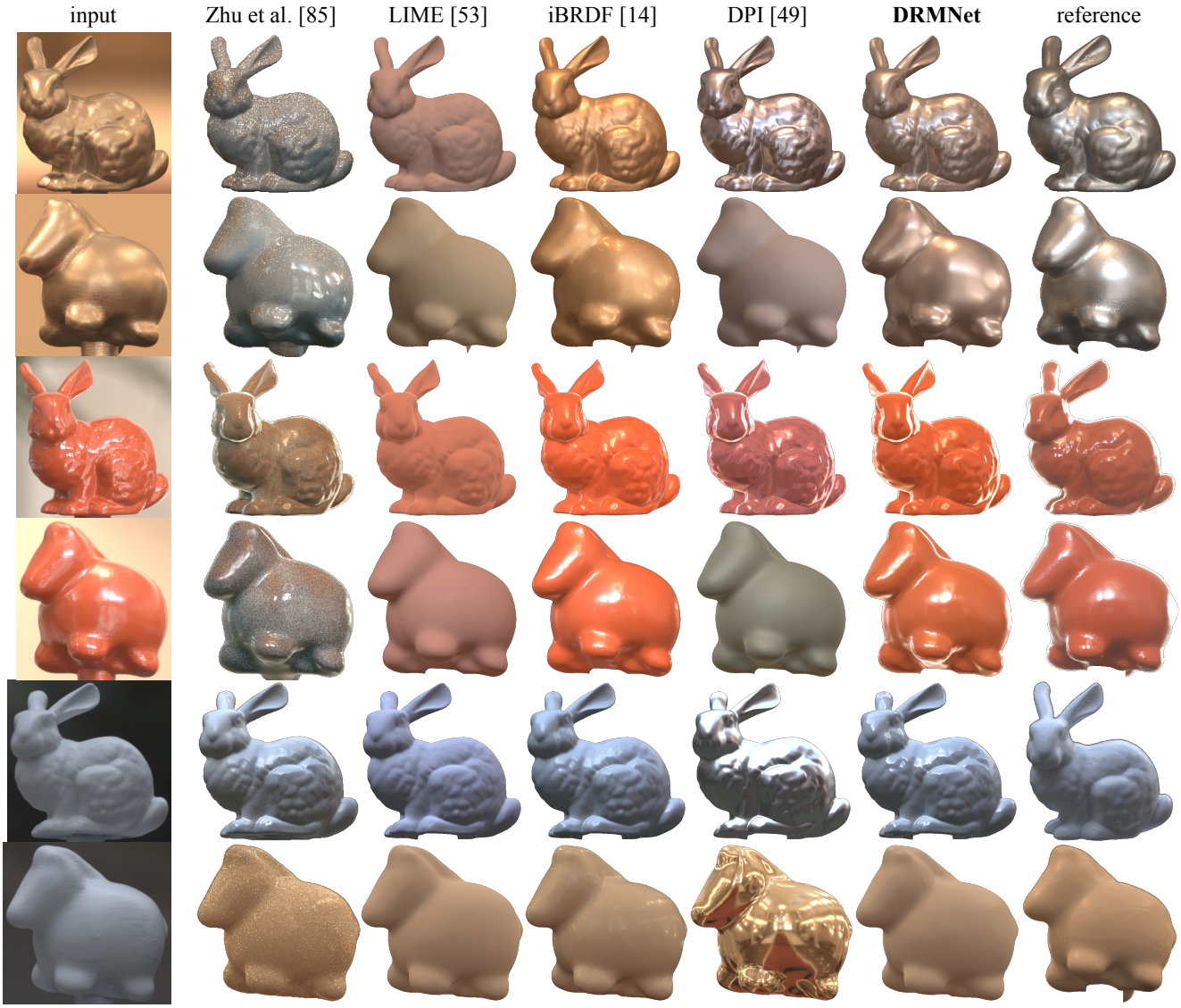

Figure 2. Additional relighting results for the nLMVS-Real dataset [75]. DRMNet achieves higher qualitative accuracy suggesting its superior accuracy of reflectance estimates.

filled with noise through concatenation to the input noise at each step. Inspired by latent diffusion [87], we train for 1000 time steps with conditioned diffusion model and use 50 steps of DDIM [88] at inference to sample a completed observed reflectance map.

All networks are trained with exponential moving average of decay rate 0.9999.

The trajectory of the reflectance parameter $\Psi^{(k)}$ that dictates the forward and reverse process is determined by

$$\Psi^{(k-1)} - \Psi_0 = \eta(\Psi^{(k)} - \Psi_0)\,, \tag{21}$$

where η controls the rate of change towards perfect mirror reflection Ψ_0. In our experiments, we use $\eta = 0.95$. We set ϵ which is used as a threshold to determine convergence of $\Psi^{(k)}$ to Ψ_0 to 0.01. This means that the maximum time step is $K = 108$. The variance of the additive Gaussians for the forward and reverse steps are set to $\sigma = 0.02$ and $\delta = 0.025$.

D. Dataset

We create a large-scale synthetic reflectance map dataset. We use the Laval Indoor Dataset [10, 22] and the Poly Haven HDRIs [2] as the illumination. We split each of these HDR environment map datasets $8 : 2$ into training and test sets and combine them to obtain the overall training and test sets. Every time we sample an illumination from the training set, we sample a random reflectance parameter

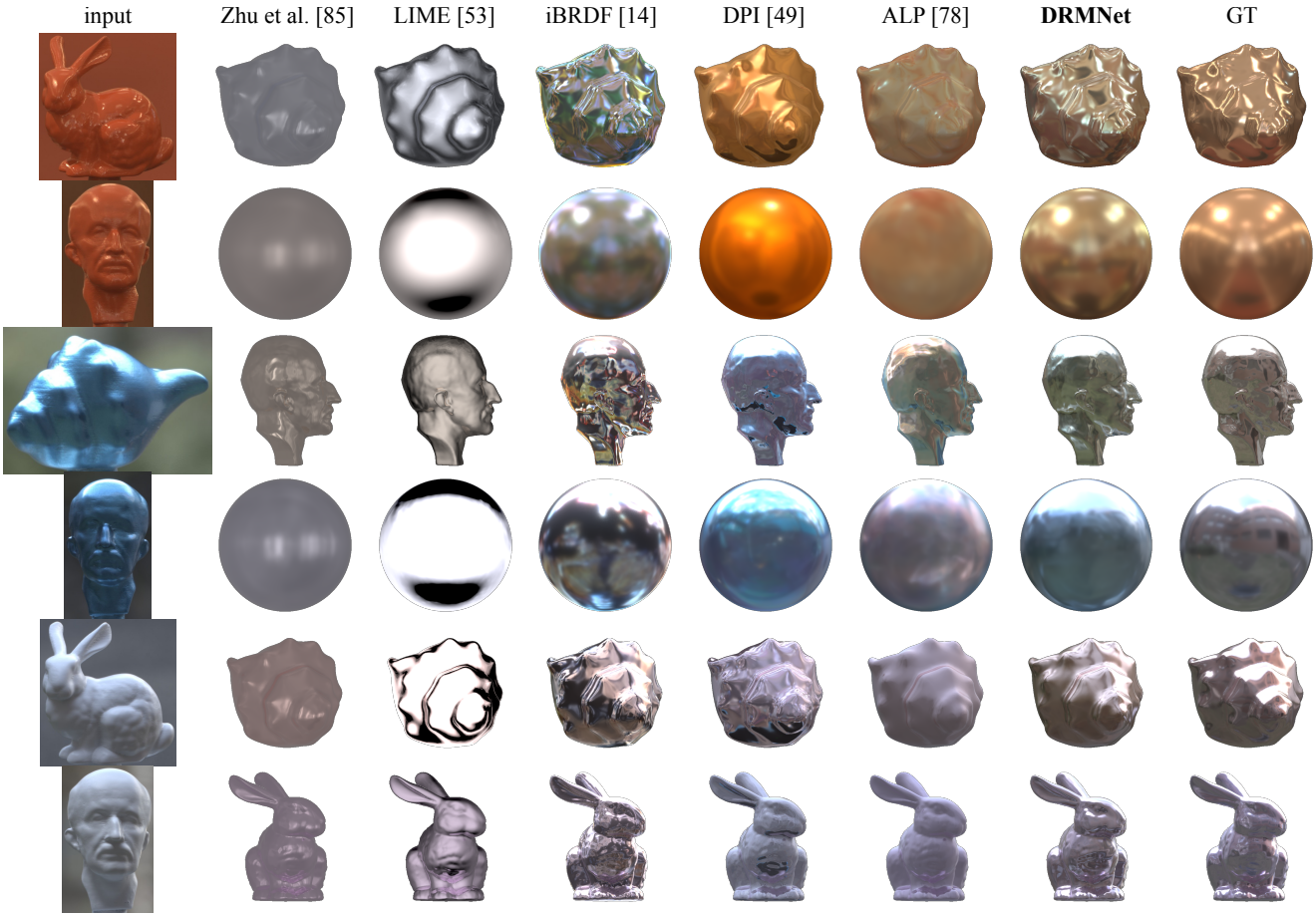

Figure 3. Additional object replacement results for the nLMVS-Real dataset [75]. DRMNet results in qualitatively higher accuracy suggesting its superior accuracy of illumination estimates.

$\Psi^{(K)}$, viewing direction, and time step k and render three reflectance maps $L_r^{(K)}$, $L_r^{(k)}$, and $L_r^{(k-1)}$ corresponding to $\Psi^{(K)}$, $\Psi^{(k)}$, and $\Psi^{(k-1)}$, respectively. The reflectance parameter $\Psi^{(K)}$ is uniformly sampled $\mathbb{R}^6; \mathbb{R} \in [0, 1]$. The view direction is sampled from a uniformly discretized set of 64 angles spanning 360 degrees, and the time step k is uniformly sampled within the range of $||\Psi^{(0)} - \Psi_0||_2 < \epsilon$. Fig. 1 shows samples of these synthetic complete observed reflectance maps. We used 1730 environment maps for training and randomly sampled reflectance parameters and viewpoints. The resulting number of reflectance maps used for training is about 1.7 million. Training took about 5 days with an NVIDIA A100 GPU.

These synthetic complete observed reflectance maps are also used to train ObsNet. We compute normal maps of the random shapes in [89] to obtain visibility masks of the reflectance map. By adding Gaussian noise to the rendered observed reflectance map and then by masking it with this visibility mask, we obtain sparse observed reflectance maps, which are paired with its complete original to train ObsNet. We also fine-tune it with raw reflectance maps from synthetically rendered random shapes for robustness against global illumination. As all reflectance maps are in HDR, we apply log-scale transformations to pass them through each network. For ObsNet, we take the logarithm for each reflectance map independently and linearly map them to $[-1, 1]$. For DRMNet, we normalize the overall scale of the forward and reverse process based on the intensity of the observed reflectance map and compress the brighter values with $\log_{10}(x + 0.1) + 1$. The Gaussian noise for each model is applied after these intensity transforms.

E. Implementations of Past Methods

In this section, we elaborate on the prerequisites and implementations of previous methods for comparative experimental evaluation.

LIME [53] is a method for estimating homogeneous reflectance and an environment map. While this method does not use the object geometry to estimate the reflectance, it

needs a normal map corresponding to the input object image for environment map estimation. We use the ground truth normal map for fairness. The environment map is estimated by mapping specular reflection using the normal map and adding approximate low frequency illumination with spherical harmonics up to the third order from diffuse reflection.

DPI [49] estimates a spatial-varying BRDF and environment map from images. For fair comparison, we constrain the BRDF to be homogeneous. Otherwise, the surrounding environment reflected on the object surface would be baked into the spatial-varying BRDF and the estimated environment maps become random.

ALP [78] needs to first compute a spatial-varying BRDF from multiple images of an object with known illumination and geometry. Only after that, the method can estimate an environment map from an image of the same object placed in a different environment. To compare this method with ours on the nLMVS-Real dataset, we use the multi-view images in the "laboratory" environment to pre-acquire the BRDF. We use this pre-acquired BRDF to run ALP on images in other environments, *i.e*., "buildings/chapel/court/entrance/manor."

Zhu *et al*. [85] estimate spatial-varying BRDF, geometry, and out-of-view illumination from a single image of complex indoor scenes. We obtain complete environment maps by estimating the out-of-view area at the center position of the input image from the network. This method is significantly different from ours in its assumptions (wide field-of-view input images), so we only compare it within this supplemental material.

F. Additional Qualitative Results

Figure 2 and Fig. 3 show additional relighting and object replacement results for the nLMVS-Real dataset [75]. We also compare with Zhu *et al*. [85] as it explicitly recovers the BRDF and illumination in the course of indoor inverse-rendering. As it estimates spatial-varying BRDF, the results of each method in Fig. 2 show the objects with the same orientation as the inputs. On the other hand, the objects in the last column labeled as "reference" have a different orientation, because the viewpoint varies across environments in the nLMVS-Real dataset. Figure 4 shows the quantitative results on the Delight-Net dataset set [27]. Our results are qualitatively more accurate, cleanly recovering the missing frequency components of the illumination, also evident in the object replacement results, while attaining more accuracy reflectance close to the ground truth relighting compared with other methods which reconstruct arbitrary frequency characteristics of the illumination. iBRDF [14] comes close in quantitative accuracy but the high-frequency components of the illumination tends to be overestimated as evident in the object replacement results. Note again that ALP knows the reflectance.

G. Ablation Study

We validate the architecture of DRMNet by ablating its components and comparing them with the full model. We consider three sets of ablation studies. "W/o $\Psi^{(k)}$" eliminates the step-wise reflectance estimate as input to IllNet so that the illumination and reflectance estimation are achieved independently. This ablation studies the importance of the confluence of jointly estimating the reflectance and the illumination, rather than independently. "W/o $L_r^{(K)}$" eliminates the conditioning on the observed reflectance map $L^{(K)}$ and achieves the iterative inversion solely based on the previous reflectance map estimate. This ablation studies the importance of referring to the observed reflectance map at every step of illumination estimation. "Once" estimates the reflectance from the observed reflectance map $L_r^{(K)}$ and reuses this initial estimate in the recursive diffusion process.

Table 1 shows quantitative results. The results clearly show that conditioning IllNet on the observed reflectance map $L_r^{(K)}$ which explicitly embodies the forward radiometric forward process is essential, and that the illumination and reflectance estimation processes are intertwined so that conditioning IllNet on the current estimate and recursively refining the reflectance estimate itself leads to more accurate estimates of both. This is likely because the estimation in the DRMNet recursion operates like alternating optimization leading to stable and consistent estimation. The conditioning on $L_r^{(K)}$ helps the recursive estimation of the illumination and reflectance to remain consistent with the object appearance when combined acting like a reconstruction loss.

In contrast, the ablation result on the reflectance estimation "once" is counter-intuitive and we find it to be inconclusive. Estimation of the reflectance in one-shot leads to higher accuracy in logRMSE of the reflectance estimate. The reason why we still employ the iterative reflectance estimation is that we empirically found that the training and inference were more robust with this choice (also seen in the slight drop in logRMSE of illumination estimates), especially for real data. We believe this discrepancy from intuition and empirical test, manifesting particularly in logRMSE also reflects the difficulty of evaluating the "goodness" of a network for a generative task. We plan to further study this in more detail.

H. Stochastic Behavior

We analyze the stochastic variability of our method. DRMNet seamlessly integrates stochasticity in the inverse rendering process via a reverse diffusion process on the additive Gaussian observation noise of radiometric image for-

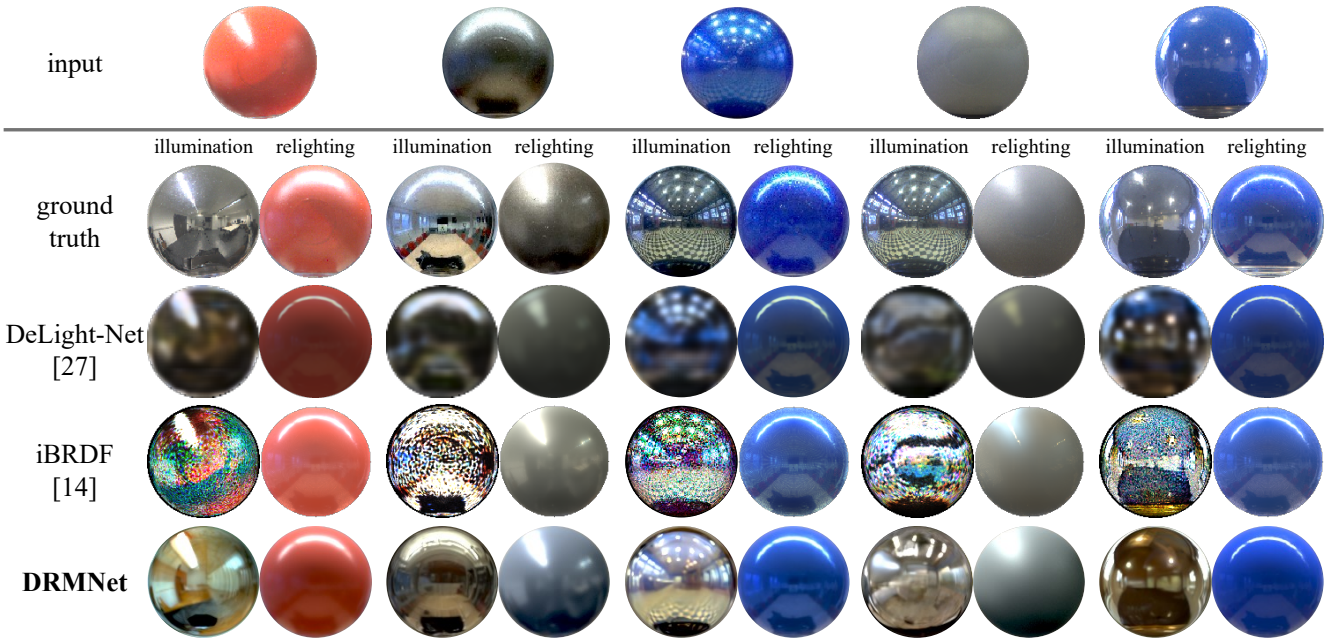

Figure 4. Qualitative results for the DeLight-Net dataset set [27]. In comparison with iBRDF [14] and DeLight-Net [27], DRMNet achieves qualitatively natural estimation for illumination and reflectance.

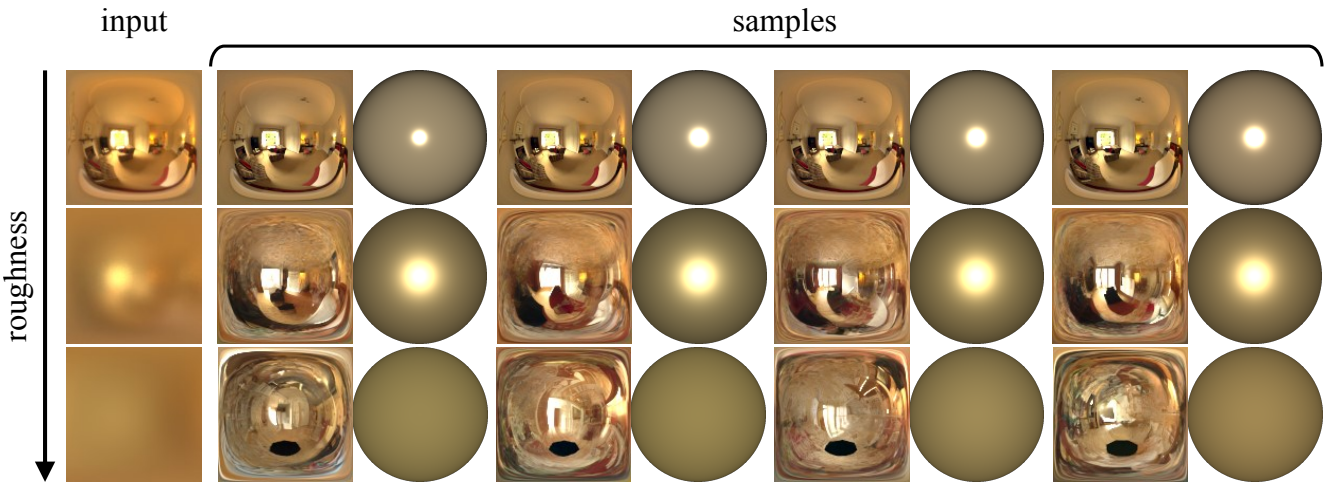

Figure 5. Results of multiple runs of DRMNet on the same input image. The left column shows the observed reflectance maps, and the right columns show the estimated samples of illumination (left) and reflectance (right) for each observation. See text for details.

mation. This enables estimation of illumination faithful to the observation with stochastic variability without separate sampling. Figure 5 shows the results of estimating the illumination and reflectance multiple times for the same input images from a set of input images under different illumination and of objects with different surface roughnesses. For the same observed reflectance map, a variation of illumination environments are estimated and their variance is large for dull reflectance closer to Lambertian and decreases for more specular reflectance centered around the ground truth. The larger the surface roughness, the wider-band of high-frequency of illumination are attenuated which is accurately reflected in these results. Note how well the recovered reflectance maps preserve the overall structure of the illumination up to the necessary frequencies—it respects the observation as much as it needs to. This is in sharp contrast to other methods that completely hallucinate an environment from noise [49]. The reflectance estimates vary accordingly which are consistent with the observed reflectance map when combined with the corresponding illumination

	illumination			reflectance
	logRMSE↓	SSIM↑	LPIPS↓	logRMSE↓
w/o $\Psi^{(k)}$	2.87	0.42	0.54	0.29
w/o $L_{\mathrm{r}}^{(K)}$	2.50	0.40	0.57	<u>0.25</u>
once	<u>2.45</u>	**0.46**	**0.51**	**0.21**
full model	**2.41**	**0.46**	**0.51**	<u>0.25</u>

Table 1. Ablation studies of DRMNet. The full model achieves the highest accuracy, confirming the importance of principled joint estimation of illumination and reflectance with reference to the observed reflectance map realized through interdependent conditioning within DRMNet.

estimates. These results clearly show that DRMNet canonically solves stochastic inverse rendering while capturing the ambiguity between the illumination and reflectance.

References

[87] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[88] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 3

[89] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep Image-Based Relighting from Optimal Sparse Samples. *ACM TOG*, 37(4):126, 2018. 4